Location, Location, Location: Using Central Nodes for Efficient Data Collection in WSNs

Vitaly Milyeykovski*, Michael Segal, *Senior Member, IEEE**, Hanan Shpungin†
* Department of Communication Systems Engineering Ben-Gurion University of the Negev
Beer-Sheva, 84105, Israel, e-mail: milyeykovski@gmail.com, segal@cse.bgu.ac.il
† Department of Electrical Computer Engineering University of Waterloo
Waterloo, N2L 3G1, Canada, e-mail: hanan.shpungin@uwaterloo.ca

***Abstract*—We study the problem of data collection in Wireless Sensor Networks (WSN). A typical WSN is composed of wireless sensor nodes that periodically sense data and forward it to the base station in a multi-hop fashion. We are interested in designing an efficient data collection tree routing, focusing on three optimization objectives: energy efficiency, transport capacity, and hop-diameter (delay).**

In this paper we develop single- and multi-hop data collection, which are based on two definitions of node centrality: centroids and balance nodes. We provide theoretical performance analysis for both approaches, present their distributed implementation and discuss the different aspects of using each. Most of our results are for two-dimensional WSNs, however we also show that the centroid-based approach is asymptotically optimal in three-dimensional random node deployments. We also show several simulation results that support our theoretical findings.

I. Introduction

A wireless sensor network (WSN) consists of small autonomous low-cost low-power devices that carry out monitoring tasks. Initially developed for military use, WSNs can nowadays be found in many civil applications, such as environmental monitoring, biomedical research, seismic monitoring, and precision agriculture [1]. The devices are called *sensor nodes* and the monitored data is typically collected at a *base station*, following a specific collection pattern of activated wireless links.

As these networks have no hard-wired underlying topology, one of the most fundamental issues when a WSN is deployed is the formation of an efficient communication backbone, or in other words, answering the question *which links to use in order to collect the data from the sensor nodes?*

Efficiency can be defined in many ways, for example it can be maximizing the rate at which data is collected ([19], [41], [43]) from the sensor nodes, prolonging the network lifetime by reducing the energy consumption ([5], [8], [30], [32], [36]), minimizing the number of hops from the sensor nodes to the collecting base station ([13], [18]), and other optimization objectives. It is apparent that the *topological structure of the communication backbone* plays a vital role in its efficiency. However, it is also important to note that a communication backbone which has good performance in some of the criteria can have a bad one in others. For example, using the minimum spanning tree (MST) as the backbone provides an optimal network lifetime performance for same initial battery charges [4], however it can have a very poor hop-diameter, which is critical for delay minimization.[1] Thus, the network designer has to take special care when deciding which links to activate for the purpose of data collection, as different optimization objectives may be have a negative effect on each other.

The problem of data collection can be divided into two major paradigms. Data collection *with aggregation* ([21], [39]) allows each sensor node to accumulate the messages of its descendants and then pass only one fixed-size message towards the base station. The second paradigm, is data gathering *without aggregation* ([25], [26]) which requires that *all* messages initiated by the sensors will eventually reach the base station.

Our main objective in this paper is to construct efficient communication backbones for single- and multi-hop data collection with aggregation in WSNs for both random and arbitrary sensor node deployments, while measuring the efficiency based on the next three metrics.

- The *transport capacity* metric represents the sum of rate-distance products over all the active links. It is measured in *bit-meters* and was first introduced by [16]. The idea behind this measure is to capture both the notion of the overall rate and distance that the information travels in a network.
- *Hop-diameter* is another important metric which reflects the depth of the data gathering tree, i.e. the maximum number of hops from any of the sensor nodes to the base station.
- *Total energy consumption* is probably one of the most important parameters of a WSN as the sensor nodes are often deployed in areas where battery replacement is infeasible [7]. Wireless communication is a major contributer to the energy budget of a node. In this paper we focus on minimize the total energy consumed by all nodes for communication purposes.

We propose a novel approach for the construction of

[1] Imagine n sensor nodes located on a straight horizontal segment, with the base station being to the right of the right-most sensor. It is easy to show that the hop-diameter of MST in this case is n.

communication backbones by identifying *central locations* in the deployment area and routing all data through these regions. The general idea is that these locations would serve as aggregation points both on a local and global scale. In particular, we use an interesting geometrical notion of *centroids*, which is defined as the central geometrical position of a collection of nodes, which are used as a guide for the construction of hierarchical aggregation trees. In addition, we also examine *balance nodes*, where the main motivation is to build data collection routes based on centrally located nodes in topologies which are already efficient in terms of some of the metrics.

The rest of this paper is organized as follows. In Section II we present our system settings and state our objective. Related works are surveyed in Section III. Sections IV and V are the technical sections of the paper and show the construction of data collection communication backbones for three scenarios, single-hop general network and multiple-hop random network. Simulation results for random networks are presented in Section VI. Finally, we conclude and discuss future work in Section VII.

II. System settings

A WSN consists of n wireless sensor nodes, $S = \{s_1, \ldots, s_n\}$, distributed in some area A. These nodes perform monitoring tasks and periodically report to a base station r which is located somewhere within the area A (we consider different locations throughout the paper). During the report phase, the sensor nodes propagate a message to the base station through a *data collection tree*, $T_S = (S \cup \{r\}, E_S)$, rooted at r. We consider *data collection with aggregation*, where every node $s \in S$ forwards a single unit size *report message* to its parent. The message holds an accumulated information collected from a subtree of T_S rooted at s. An example of this scenario can be found in temperature monitoring systems for fire prevention, intrusion detection, seismic readings, etc.

We assume the use of *frame-based* MAC protocols which divide the time into frames, containing a fixed number of slots. The main difference from the classic TDMA is that instead of having one access point which controls transmission slot assignments, there is a localized distributed protocol mimicking the behavior of TDMA. The advantage of a frame-based (TDMA-like) approach compared to the traditional IEEE 802.11 (CSMA/CA) protocol for a Wireless LAN is that collisions do not occur, and that idle listening and overhearing can be drastically reduced. When scheduling communication links, that is, specifying the sender-receiver pair per slot, nodes only need to listen to those slots in which they are the intended receiver – eliminating all overhearing. When scheduling senders only, nodes must listen in to all occupied slots, but can still avoid most overhearing by shutting down the radio after the MAC (slot) header has been received. In both variants (link and sender-based scheduling) idle listening can be reduced to a simple check if the slot is used or not. Several MAC protocols were developed to adapt classical TDMA solutions which use access points to ad-hoc settings that have no infrastructure; these protocols employ a distributed slot-selection mechanism that self-organizes a multi-hop network into a conflict-free schedule (see [31], [42]).

Let $d(u, v)$ be the Euclidean distance between two sensor nodes $u, v \in S$. It is customary to estimate that the energy required to transmit from u to v is proportional to $d(u, v)^\alpha$, where α is the *path-loss coefficient*. In perfect conditions $\alpha = 2$, however in more realistic settings (in presence of obstructions or noisy environment) it can have a value between 2 and 4 (see [29]). In this paper we assume $\alpha = 2$ for simplicity. However, it is possible to extend our results for other values of α which are greater than 2.

Let $E(T_S)$ be the **energy requirement** to execute a single report phase. Note that every sensor performs a single transmission, during which it sends a single message to its parent in T_S. Thus, the energy requirement is proportional to the sum of squares of lengths of the edges E_S. The focus of this paper is to study the asymptotic performance of data collection trees, thus we can express $E(T_S)$ as follows, $E(T_S) = \sum_{(u,v)\in E_S} d(u, v)^2$.

Minimizing the energy requirement is one of the primary optimization objectives when deploying a WSN due to the very low battery reserves at the sensor nodes and the high costs that are associated with replacing these batteries (if at all possible).

Another critical aspect in the design of a WSN is the **hop-diameter** of T_S. The data flows from the leafs of the delivery tree to the base station, where each intermediate node waits to receive the report messages from all its children, before sending its own report message to its parent. Therefore, the hop-diameter of T_S, denoted as $H(T_S)$, determines the delay of data collection.

The third measure that we are interested in is **transport capacity**, $D(T_S)$, of the data collection tree T_S. As mentioned earlier, the main idea which stands behind this metric is to capture the spatial rate of the network, which is represented by the total rate over some distance. In our scenario, the rate on all links is fixed as all the nodes transmit an aggregated, unit-size message, to the parent in the collection tree and the schedule is conflict-free. Thus, to maximize the transport capacity we need to minimize the total distance traveled by information, which is the sum of lengths of all the links, $D(T_S) = \sum_{(u,v)\in E_S} d(u, v)$.

Unfortunately it is impossible to achieve optimal performance in all three measures at the same time. For example, minimizing the hop-diameter results in all nodes transmitting to the base station, which is disastrous in terms of energy consumption, whereas the best topology to minimize energy consumption[2] results in a relatively high hop-diameter.

Our main objective in this paper is to construct data

[2]As described later in the paper, using the Euclidean minimum spanning tree minimizes the energy consumption.

collection trees for several node distribution scenarios which produce good performance in all three measures simultaneously.

III. Related work

To the best of our knowledge, this is the first work which takes into account all of the above 3 performance measures simultaneously. Below we discuss some of the related work on data collection, energy efficiency, bounded-hop communication, and transport capacity.

In terms of **total energy consumption** measure, it was proved in [35] that using the minimum spanning tree for data collection (gathering) with aggregation achieves optimality. A different criterion used to measure energy efficiency is network lifetime, which is defined as the time the first node depletes all its power reserves due to periodic data transmission. Segal [34] developed an optimal maximum lifetime algorithm for data collection with aggregation. One of the possible variants is to allow the use of different collection trees, which makes the maximum network lifetime data gathering problem more challenging. Interestingly, if aggregation is allowed, the problem is still polynomially solvable [21], [28], and is NP-complete otherwise [26]. Kalpakis et. al. [21] developed an optimal data collection with aggregation algorithm in $O(n^{15}\log n)$ time. To counter the slowness of the algorithm, Stanford and Tongngam [39] proposed a $(1-\varepsilon)$-approximation in $O(n^3\frac{1}{\varepsilon}\log_{1+\varepsilon} n)$ time based on Garg and Könemann [15]. For more details we refer the reader to a recent survey by Ramanan et al. [23], which covers a diverse set of data gathering algorithms in ad-hoc networks.

The notion of **transport capacity** was introduced by Gupta and Kumar in [16]. They showed that for any layout of n wireless nodes in an area of size A, with each node being able to transmit W bits per second to a fixed range, the overall transport capacity is at most $(W\sqrt{An})$ bit-meters per second under both interference models (protocol and physical). In [20] the authors derive upper bounds on the transport capacity as a function of the geographic location of the nodes. It has also been shown that the scaling of transport capacity depends, among other factors, on channel attenuation and path loss [44], [45], [46].

Finally, communication backbones with **bounded hop-distances** between participating nodes has also been studied. For the linear layout of nodes and an upper bound on hop-distance, Kirousis et al. [22] developed an optimal power assignment algorithm for strong connectivity in $O(n^4)$ time. In the Euclidean case, [10] obtains constant ratio algorithms for the bounded-hop vertex connectivity for well spread instances. Beier et al. [3] proposed an optimal algorithm to find a bounded-hop minimum energy path between paris of nodes. In [6] the authors obtain bicreteria approximation algorithms for connectivity and broadcast while minimizing the hop-diameter and energy consumption. Funke and Laue [14] provide a PTAS for the h-broadcast algorithm in time linear in n. Additional results for bounded range assignments can be found in [9], [11], [37].

IV. Single-hop collection

We start be defining the notion of geometric centroids and then analyze the performance bounds of single-hop communication backbone which is centroid-based. In the end we discuss the possible pitfalls of using a single-hop collection tree.

For n points $P=\{p_1,p_2,\ldots,p_n\}$, $n\geq 2$, placed in the Euclidean plane, with coordinates (x_i,y_i), $i=1,\ldots,n$, and assuming general position, the *centroid* $c(P)$ is a point defined as $c(P)=(\overline{x},\overline{y})$, where $\overline{x}=\sum_{i=1}^{n}x_i/n$ and $\overline{y}=\sum_{i=1}^{n}y_i/n$, which conceptually represents the center location of P.

Apparently the centroid of n points has two very interesting properties as outlined in the following theorems that provide an analysis of the sum of squares of distances, which was done in [24], and sum of distances, which we develop here, between the points and the centroid.

Theorem 4.1 ([24]): For any set of points P and an arbitrary point p' in the Euclidean plane, $\sum_{p\in P}d(p,c(P))^2\leq\sum_{p\in P}d(p,p')^2$.

Theorem 4.2: For any set of points P and an arbitrary point p' in the Euclidean plane, $\sum_{p\in P}d(p,c(P))\leq 2\sum_{p\in P}d(p,p')$.

Proof: Let p^* be the geometric median[3] of points P. Clearly for every $p\in P$, $d(p,c(P))\leq d(p,p^*)+d(p^*,c(P))$, and thus $\sum_{p\in P}d(p,c(P))\leq\sum_{p\in P}d(p,p^*)+|P|\cdot d(p^*,c(P))$. From the convexity of the Euclidean norm it follows that the norm of an average of a set of points is at most the average of the norms of the points in the set, that is $d(p^*,c(P))\leq\sum_{p\in P}d(p,p^*)/|P|$. Therefore for any $p'\in R^2$, $d(p^*,c(P))\leq 2\sum_{p\in P}d(p,p^*)\leq 2\sum_{p\in P}d(p,p')$. ∎

Clearly, the bounds shown above represent the total energy consumption and transport capacity measure for a single-hop data collection tree if the base station is located at the centroid of the sensors, which is reasonable to expect.

Unfortunately, if we consider a single-hop tree T_S rooted at the centroid and spanning all the nodes it may be inefficient in terms of energy consumption and transport capacity. Consider the linear layout of nodes as depicted at Figure 1. If we consider the minimum spanning tree (MST) of the nodes, we obtain $D(MST)=E(MST)=n-1$. On the other hand, for T_S, we have $D(T_S)=\Omega(n^2)$ and $E(T_S)=\Omega(n^3)$, which is quite poor compared to the one obtained by the MST.

In the next section we describe a hierarchical topological structure, the *k-layer centroid network*, which is used for multi-hop data collection in random WSNs and achieves better performance than the single-hop tree.

[3]The geometric median p^* of a point set P is the point in the Euclidean plane that minimizes the sum of distances between itself and the points in P, i.e. $\forall p':\sum_{p\in P}d(p,p^*)\leq\sum_{p\in P}d(p,p')$.

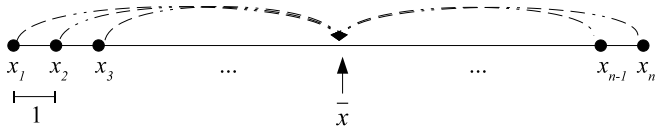


Fig. 1. Worst case performance of single-hop centroid-based data collection tree.

V. Multi-hop collection for random deployments

Our first construction of multi-hop data collection is for randomly deployed sensor networks. In this scenario we assume that n sensor nodes are randomly and independently placed in the area A with uniform distribution. We also assume that A is a unit square. We show an efficient communication backbone construction which is based on *centroid networks*, which are hierarchical geometrical structures on top of a point set P which represents the sensor nodes. As in the single-hop scenario, we assume the base station is located at $c(P)$.

1) k-layer centroid networks: We start by providing the definitions and notation used in the context of k-layer centroid networks and then proceed to presenting several useful properties and observations regarding these networks.

The *k-layer centroid network, $k > 2$, based on a point set P* (in short, k-centroid network), is a k-layer undirected tree $T_P = (V, E)$, where V and E are the node and edge sets, respectively. The leafs of the tree $V_P \subset V$ represent the points P, and the internal nodes $V_C = V \setminus V_P$ represent the centroids of subsets of P. Let r be the root of the tree. For convenience we use the notion of *node* and *point* interchangeably instead of saying *node that represents a point*.

The nodes V are divided into k layers, $V_1, \ldots, V_k$ such that $V_1 = \{r\}$, $V_C = V_1 \cup \ldots \cup V_{k-1}$, and $V_k = V_P$. The edges E connect between nodes in adjacent layers such that the parent of $u \in V_i$, $\pi(u)$, is in V_{i-1} and the children of $v \in V_j$, $N(v)$, are in V_{j+1}, for any $i, j, 1 < i \leq k, 1 \leq j < k$. We use E_i to denote the set of edges between layers i and $i+1$, $1 \leq i \leq k-1$. In a k-centroid network the following two conditions hold for any node $v \in V_C$:

- $|N(v)| > 0$.
- Let T_v be a subtree of T, rooted at $v \in V_C$, and let P_v be the points represented by the leafs of T_v. Then, v is the centroid of P_v.

For example, Fig. 2 shows a 3-centroid network where the second layer nodes are centroids of points sets P_1, P_2, P_3, P_4.

Note that according to the second condition above, the root of the tree is the centroid of the whole point set P. Next we provide several useful properties of k-centroid networks. We start with an observation which follows directly from the definition above.

Observation 5.1: Let T_P be a k-centroid network, $k > 2$, and let v be a non-leaf node in T_P with a height of l. Then T_v is an $(l+1)$-centroid network based on P_v and if $l > 1$ then for every $u \in N(v)$, T_u is an l-centroid network based on P_u.

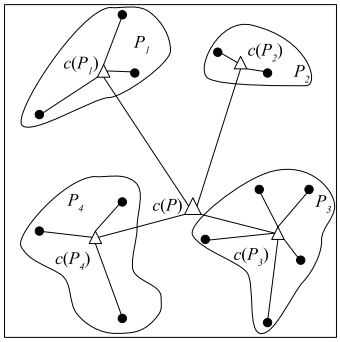


Fig. 2. An example of a k-centroid network.

Another interesting characteristic of k-centroid networks is that it is possible to easily add or remove layers with only local changes to the edge set E. We refer to the process of adding layers as *extension* and to the removal of layers as *simplification*. Let T_P be the original k-centroid network. A **simplified network** $T_P^{(-i)} = (V^{(-i)}, E^{(-i)})$ is obtained by removing the i-th layer, $1 < i < k$ (the root and the leafs cannot be removed), and connecting the parent of every removed node to its grandchildren in T_P. Formally,

$$V^{(-i)} = V \setminus V_i$$

and

$$E^{(-i)} = (E \setminus E_i) \cup \{(\pi(u), v) : u \in V_i, v \in N(u)\}.$$

Adding a layer to **extend the network** is essentially providing an additional level of grouping the points into subsets. To add a layer below an existing i-th layer, $1 \leq i < k$ (it is not possible to extend the network below the leafs layer), we need to remove the edges that connect layers V_i and V_{i+1}, and to add new edges which connect the new layer to the rest of the tree. Formally, for a k-centroid network, the new $(k+1)$-centroid network, $T^{(+i)} = (V^{(+i)}, E^{(+i)})$, is defined as follows. For each node $u_j \in V_i$, $1 \leq j \leq |V_i|$, we partition its children $N(u_j)$ into m_j, $1 \leq m_j < |N(u_j)|$ disjoint subsets $U_1^j, \ldots, U_{m_j}^j \subseteq N(u_j)$. Then, the new nodes of the added layer, $V_{[i] \leftrightarrow [i+1]}$, are the centroids of the union of the leafs in the trees rooted at the nodes of these subsets, that is

$$V_{[i] \leftrightarrow [i+1]} = \{u_l^j : 1 \leq j \leq |V_i|, 1 \leq l \leq m_j\},$$

where

$$u_l^j = c(\{p : p \in P_v, v \in U_l^j\}).$$

The edge set is modified by disconnecting V_i and V_{i+1} and connecting these layers to the new nodes,

$$\begin{aligned} E^{(+i)} =& (E \setminus E_i) \cup \{(u_j, u_l^j) : 1 \leq j \leq |V_i|, 1 \leq l \leq m_j\} \\ & \cup \{(u_l^j, v) : 1 \leq j \leq |V_i|, 1 \leq l \leq m_j, v \in U_l^j\}. \end{aligned}$$

It is easy to see that in both cases, the result of either simplification or extension is a proper $(k-1)$- or $(k+1)$-centroid network, respectively. Also note that the simplification process is deterministic for every removed layer, whereas in the case of extension there are multiple possible outcomes.

The following theorem shows that by extending an existing k-centroid network we actually reduce the sum of squares of distances in the network. For the ease of exposition we denote by $d(u,v)$ the distance between the points or centroids represented by the nodes $u, v \in V$, and by $S_2(T_P)$ the sum of squares of distances in T_P, i.e. $S_2(T_P) = \sum_{(u,v)\in E} d(u,v)^2$.

Theorem 5.2: Let T_P be a k-centroid network based on a point set P. For any l-centroid network, T'_P, which can be obtained through a series of extensions from T_P, it holds that $S_2(T_P) \geq S_2(T'_P)$.

Proof: In order to prove the theorem, it is enough to show that for $l = k+1$ it holds $S_2(T_P) \geq S_2(T'_P)$. Let T'_P be obtained from T_P by adding a layer below some existing i-th layer, for some i, $1 \leq i < k$. We are going to show, separately for each centroid $u_j \in V_i$, $1 \leq j \leq |V_i|$, that $\Sigma_{(x_m,y_m)\in N(u_j)} d(u_j,(x_m,y_m))^2 \geq \Sigma_{l=1}^{m_j} d(u_j,u_l^j)^2 + \Sigma_{l=1}^{m_j}\Sigma_{p\in U_l^j} d(u_l^j,p)^2$. Subtract from the left side of the inequality the right side. If the theorem is true, the result should be non-negative.
$\Sigma_{(x_m,y_m)\in N(u_j)} d(u_j,(x_m,y_m))^2 - \Sigma_{l=1}^{m_j} d(u_j,u_l^j)^2 - \Sigma_{l=1}^{m_j}\Sigma_{p\in U_l^j} d(u_l^j,p)^2 = \big\langle \Sigma_{l=1}^{m_j}|U_l^j| x_{u_j}^2 - 2\Sigma_{l=1}^{m_j}|U_l^j| x_{u_j} x_{u_l^j} + \Sigma_{l=1}^{m_j}|U_l^j| {x_{u_l^j}}^2 - \Sigma_{l=1}^{m_j} x_{u_j}^2 + 2\Sigma_{l=1}^{m_j} x_{u_j} x_{u_l^j} - \Sigma_{l=1}^{m_j}|U_l^j| {x_{u_l^j}}^2 \big\rangle + \big\langle \Sigma_{l=1}^{m_j}|U_l^j| y_{u_j}^2 - 2\Sigma_{l=1}^{m_j}|U_l^j| y_{u_j} y_{u_l^j} + \Sigma_{l=1}^{m_j}|U_l^j| {y_{u_l^j}}^2 - \Sigma_{l=1}^{m_j} y_{u_j}^2 + 2\Sigma_{l=1}^{m_j} y_{u_j} y_{u_l^j} - \Sigma_{l=1}^{m_j}|U_l^j| {y_{u_l^j}}^2 \big\rangle = \Sigma_{l=1}^{m_j}(|U_l^j| - 1)(x_{u_j} - x_{u_l^j})^2 + \Sigma_{l=1}^{m_j}(|U_l^j|-1)(y_{u_j} - y_{u_l^j})^2 = \Sigma_{l=1}^{m_j}(|U_l^j| - 1)((x_{u_j} - x_{u_l^j})^2 + (y_{u_j} - y_{u_l^j})^2) = \Sigma_{l=1}^{m_j}(|U_l^j|-1) d(u_j,u_l^j)^2.$
Clearly the last expression is equal or larger than 0. ■

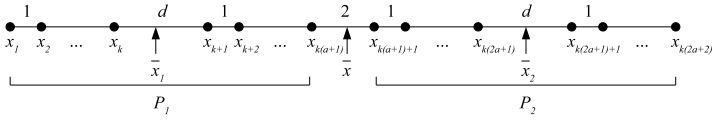


Fig. 3. Extending a k-network does not always improve the sum of distances.

Unfortunately, we cannot make a similar claim for sum of distances (as demonstrated in Fig. 3). Let us consider the set P of $k(2a+2)$, $k, a \in \mathbb{N}$, points on line, ordered in increasing order by their coordinates x_i, $i = 1, ..., n$. The distance between points x_i and x_{i+1}, for $i = 1, ..., k(2a+2)$, $i \neq k$, $i \neq k(a+1)$, $i \neq k(2a+1)$, is 1; for $i = k(a+1)$ it is 2; and for $i = k$, $i = k(2a+1)$, it is d. The coordinate of the centroid C of the points in P, is $\overline{x} = k(a+1) - 1 + d$. We partition P into two sets: $P_1 = \{x_1, ..., x_{k(a+1)}\}$ and P_2 - the remaining points. Coordinate of the centroid C_1 of the points in P_1, is $\overline{x}_1 = \frac{k(a+1)^2+2ad-3a-1}{2(a+1)}$. The sum of the distances between C and the points in P_1 is $\sum_{i=1}^{k(a+1)} (\overline{x} - x_i) = \frac{k^2(a+1)^2+k(a-1)+2kd}{2}$. The sum of the distances between C_1 and the points in S_1 is $\sum_{i=1}^{k(a+1)} |\overline{x}_1 - x_i| = \frac{k^2a(a+1)-2ak+2akd}{a+1}$. When comparing these two sums one can see that in the case of $a > 1$ and for any given k there is such d for which $\sum_{i=1}^{k(a+1)} |\overline{x}_1 - x_i| > \sum_{i=1}^{k(a+1)} (\overline{x} - x_i)$. The same happens with a symmetric case while considering the sum of distances between C and C_2 (the centroid of P_2) and the points of P_2. Thus, we obtain that the sum of distances from C to the entire set is less than the sum of distances from C_1 to P_1 plus the sum of distances from C_2 to P_2.

A. Data collection using centroid networks

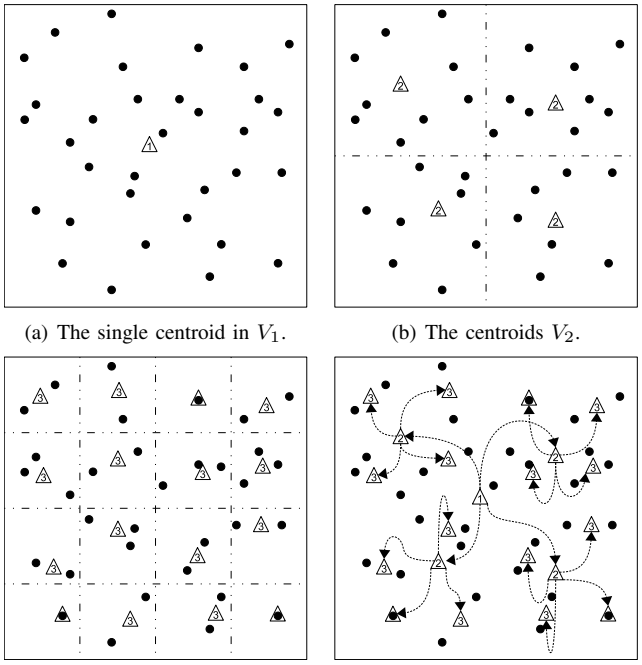
First we describe the k-centroid network $T = (V, E)$ which we then use to produce the communication backbone.

Let P be the points that correspond to the location of sensor nodes S. To construct T we repeatedly divide the unit square area A into sub-areas. First we divide A into 4 equal square sub-areas, then each of these sub-areas is further divided into 4 sub-areas, and so forth. In every sub-area we pick one centroid of the points in that sub-area and add it to T. The connections between these centroids are added according to the point hierarchy as described above, while the root of T is the centroid of all the points P. The iteration proceeds in steps, where at each step we handle subdivision of sub-areas of the same size; it ends once it is not possible to continue subdividing the areas into non-empty square regions. In the final phase, the centroids are connected to the points in their respective areas. We now describe this process in detail.

1) Let $j \leftarrow 1$, $\mathcal{A}_1 \leftarrow \{A\}$, $V \leftarrow V_1 \leftarrow \{c(P)\}$, $E \leftarrow \emptyset$.
2) While it is possible to divide all the areas in $\mathcal{A}_j$ into 4 non-empty equal square sub-areas:
   a) Initialize $\mathcal{A}_{j+1} \leftarrow \emptyset$, $V_{j+1} \leftarrow \emptyset$, $E_j \leftarrow \emptyset$.
   b) For every $A' \in \mathcal{A}_j$:
      i) Let P' be the points which are within area A'.
      ii) Divide A' into 4 equal square sub-areas A'_1, A'_2, A'_3, A'_4. Let P'_1, P'_2, P'_3, P'_4 be the points sets in these areas, respectively.
      iii) Add the centroids $c(P'_1)$, $c(P'_1)$, $c(P'_1)$, and $c(P'_1)$ to V_{j+1}.
      iv) Add the edges $(c(P'), c(P'_1))$, $(c(P'), c(P'_2))$, $(c(P'), c(P'_3))$, and $(c(P'), c(P'_4))$ to E_j.
      v) Add the areas A'_1, A'_2, A'_3, and A'_4 to $\mathcal{A}_{j+1}$.
   c) Update $V \leftarrow V \cup V_{j+1}$, $E \leftarrow E \cup E_j$, and increase $j \leftarrow j+1$.
3) Initialize $V_{j+1} \leftarrow \emptyset$ and $E_j \leftarrow \emptyset$.
4) For every $A' \in A_j$:
   a) Let P' be the points inside A'.
   b) Add all the edges $\{(c(P'), p) : p \in P'\}$ to E_j and P' to V_{j+1}.
5) Update $V \leftarrow V \cup V_{j+1}$ and $E \leftarrow E \cup E_j$.

For example of an execution, see Fig. 4.

Let k be the last value of j in the above scheme. Clearly, the obtained $T = (V, E)$ is a k-centroid network. We are now



(a) The single centroid in V_1. (b) The centroids V_2. (c) The centroids V_3. (d) The 3-centroid network (without links to the points).

Fig. 4. An example of the algorithm execution, showing the construction of a 3-centroid network (the links to the points in the last layer are omitted for clarity). The triangles represent the centroids in all three layers.

ready to define the data collection tree $T_S = (S \cup \{r\}, E_S)$. The general idea is to match between the virtual nodes in V and the sensor nodes S. For every centroid $c \in V$, let P_c be the points that produced c, and let S_c be the sensor nodes that correspond to these points. Then we choose $s_c \in S_c$ to be the sensor node which is closest[4] to c, i.e. $d(s_c) = \min_{s \in S_c} d(s, c)$. Note that we might not have a tree yet, as it is possible that there are cycles and self-loops. These are easily removed by running a breadth-first search in the obtained graph, starting with the root (the node closest to the centroid of all the sensors). The resulting breadth-first tree is T_S.

In order to estimate the efficiency of the constructed data collection tree T_S we will use the following theorems.

Theorem 5.3 ([40]): The sum of the of edges of an MST in two (three) dimensional unit size square (cube) for random uniform points is $\Omega(n^{1/2})$ ($\Omega(n^{2/3})$).

Theorem 5.4 ([33], Theorem 2.2): The sum of the squares of edges of an MST in two (three) dimensional unit size square (cube) for random uniform points is $\Omega(1)$ ($\Omega(n^{1/3})$).

We claim the following.

Theorem 5.5: For the data collection tree T_S it holds that $H(T_S) = O(\log n)$, $E(T_S)$ is $O(\log n)$ times the optimal, and $D(T_S)$ is $O(1)$ times the optimal.

Proof: Suppose we run the scheme of constructing T until $j = h+1$, $h+1 \leq k$. In other words we perform step 2 of the scheme $h-1$ times and then proceed to step 3. We obtain the $h+1$-centroid network. Denote the sum of squares of lengths of edges in E_h by $\sum_h^{[2]}$ and the sum of squares of lengths of the edges in $E \setminus E_h$ by $\sum_c^{[2]}$. Denote the sum of lengths of edges in E_h by $\sum_h$ and the sum of squares of lengths of the edges in $E \setminus E_h$ by $\sum_c$. Obviously, $E(T_S) = \sum_h^{[2]} + \sum_c^{[2]}$ and $D(T_S) = \sum_h + \sum_c$. After $h-1$ times of A partition, $\mathcal{A}_h$ has 4^{h-1} square sub-areas, each of which has diagonal of length $\sqrt{\frac{2}{|\mathcal{A}_h|}} = \sqrt{\frac{2}{4^{h-1}}}$. Since, each $\mathcal{A}' \in \mathcal{A}_h$ has one centroid (and the total amount of centroids in $\mathcal{A}_h$ is $|V_h| = |\mathcal{A}_h| = 4^{h-1}$), then $\sum_h^{[2]} \leq \left(\sqrt{\frac{2}{|\mathcal{A}_h|}}\right)^2 (n - |V_h|) \leq \frac{2}{4^{h-1}}\left(n - 4^{h-1}\right) = O\left(\frac{n}{4^{h-1}}\right)$, and $\sum_t \leq \sqrt{\frac{2}{|\mathcal{A}_h|}}(n - |V_h|) \leq \sqrt{\frac{2}{4^{h-1}}}\left(n - 4^{h-1}\right) = O\left(\frac{n}{\sqrt{4^{h-1}}}\right)$. Suppose the sensors are located on the vertex points of unit size grid, with grid cells of size $\frac{1}{\sqrt{n}-1} \times \frac{1}{\sqrt{n}-1}$. By applying the scheme of construction T for this case of sensors arrangement, it is easy to see that $\sum_c^{[2]} = O(\log |\mathcal{A}_h|) = O\left(\log 4^{h-1}\right)$ and $\sum_c = O\left(\sqrt{|\mathcal{A}_h|}\right) = O\left(\sqrt{4^{h-1}}\right)$. Thus, for the case of grid, $E(T_S) = \sum_h^{[2]} + \sum_c^{[2]} = O\left(\frac{n}{4^{h-1}} + \log 4^{h-1}\right)$ and $D(T_S) = \sum_h + \sum_c = O\left(\frac{n}{\sqrt{4^{h-1}}} + \sqrt{4^{h-1}}\right)$. Returning to the stochastic case we observe that the sums $\sum_c^{[2]}$ and $\sum_c$ are maximal, if in each $\mathcal{A}' \in \mathcal{A}_h$ there is a sensor. This implies that each $\mathcal{A}' \in \mathcal{A}_i$, $i = 1, \ldots, h-1$, is not empty. Therefore, $\sum_c^{[2]}$ and $\sum_c$ are equal to those on the grid up to a constant, even if we choose the centroids arbitrary within their corresponding sub-area. Thus, $E(T_S) = \sum_h^{[2]} + \sum_c^{[2]} = O\left(\frac{n}{4^{h-1}} + \log 4^{h-1}\right)$ and $D(T_S) = \sum_h + \sum_c = O\left(\frac{n}{\sqrt{4^{h-1}}} + \sqrt{4^{h-1}}\right)$. For $h - 1 = \log_4 n$, $E(T_S) = O(\log n)$ and $D(T_S) = O(\sqrt{n})$. Note that for the random uniform distribution of sensors in A, the algorithm for constructing T stops when the area of square in $\mathcal{A}_h$ is at most $O\left(\frac{\log n}{n}\right)$ [38]. Since $|\mathcal{A}_h| \geq \frac{n}{\log n}$, it follows that $k = O\left(\log_4 \frac{n}{\log n}\right) = O(\log n)$. Thus, $E(T_S) = O(\log n)$ and $D(T_S) = O(\sqrt{n})$.

Following Theorems 5.3 and 5.4, we conclude that $E(T_S)$ is $O(\log n)$ times the optimal, and $D(T_S)$ is $O(1)$ times the optimal solution. ■

Suppose that the sensors are uniformly distributed within a unit cube. According to the scheme, similar to the above, we can build the k-centroid network, $k = O(\log n)$. Following the assumptions and arguments similar to the above, it can be shown, that $\sum_h^{[2]} = O\left(\frac{n}{\sqrt[3]{(8^{h-1})^2}}\right)$, $\sum_h = O\left(\frac{n}{\sqrt[3]{8^{h-1}}}\right)$, and $\sum_c^{[2]} = O\left(\sqrt[3]{8^{h-1}}\right)$, $\sum_c = O\left(\sqrt[3]{(8^{h-1})^2}\right)$. Thus, $E(T_S) = \sum_h^{[2]} + \sum_c^{[2]} = O\left(\frac{n}{\sqrt[3]{(8^{h-1})^2}} + \sqrt[3]{8^{h-1}}\right)$ and

[4]The distance between a sensor node and a point is the Euclidean distance between the location of the sensor node and the coordinates of the point in the Euclidean plane.

$D(T_S) = \sum_h + \sum_c = O\left(\frac{n}{\sqrt[3]{8^{h-1}}} + \sqrt[3]{(8^{h-1})^2}\right)$. For $h-1 = \log_8 n$, $E(T_S) = O(\sqrt[3]{n})$ and $D(T_S) = O\left(\sqrt[3]{n^2}\right)$. Using the results of Theorems 5.3 and 5.4, we obtain that for three-dimensional case $E(T_S)$ and $D(T_S)$ are $O(1)$ times the optimal solution.

The distributed implementation of the k-centroid network is quite straightforward once we established connectivity between the nodes and chose the leader (the root of the tree). In order to establish connectivity we can use 2 different approaches. The first, described in Dolev et al. [12] forms a temporary underlying topology in $O(n)$ time using $O(n^3)$ message. The second (better) approach is given by Halldórsson and Mitra [17] that show how to do this in $O(poly(\log\beta, \log n))$, where β is the ratio between the longest and shortest distances among nodes. After the topology is established, we can use leader-election algorithm by Awerbuch [2] that shows how to find a leader in a distributed fashion in a network with n nodes in $O(n)$ time using $O(n \log n)$ messages. Next, using the location of every node, the root of k-centroid network can be determined and the construction of the k-centroid network is started in recursive fashion that takes $O(k)$ time and $O(n)$ messages assuming omnidirectional communication.

VI. Simulation results

In this section we show some simulation results of the k-centroid network constructed for the multi-hop random scenario as described in Section V. As we show, the simulation results fully support and even slightly outperform our theoretical analysis. In what follows we compare the k-centroid network topology (**KNETW**) with the optimal one, in terms of both energy consumption and total link distance, which is achieved by using the minimum spanning tree (**MST**) as the delivery tree. The optimality of MST is straightforward in the case of total distances, whereas for energy consumption it was shown to be the best possible in [35].

In our experiments we have randomly and uniformly distributed n sensor nodes in a unit square, with the network size n ranging from 50 to 950 in steps of 50. We have computed the total distance of the communication links (Fig. 5), the energy consumption (Fig. 6), and the hop-diameter (Fig. 7) of both topologies (**KNETW** and **MST**). The results below are an average of 5 tries for every network size n.

In terms of total distance (Fig. 5), we (**KNETW**) are consistently within a factor of 2.2-2.5 from the best possible (**MST**), which matches the theoretical result of the $O(1)$ approximation ratio.

Interestingly, the energy consumption of our scheme (Fig. 6) slightly outperforms the projected theoretical bound of $O(\log n)$, with the ratio rising from $3-4$ for smaller networks and up to 7 for larger ones with $n \geq 800$.

Finally, the hop-diameter of our scheme is very close to the optimum (which is obviously 1), being only 4 for a 950-node WSN. We have used a logarithmic scale in (Fig. 7) to compare it to the one produced by **MST**, which is $10-15$ times greater than **KNETW** for small networks (n from 50 to 250), and as high as 30 times greater for larger ones ($n \geq 750$).

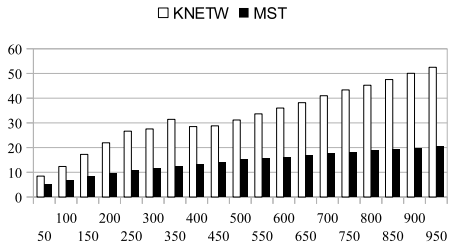


Fig. 5. Total link distances, $D(\cdot)$.

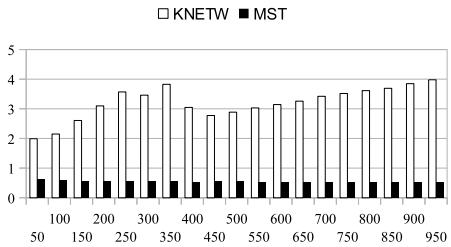


Fig. 6. Total energy consumption (sum of square of distances), $E(\cdot)$.

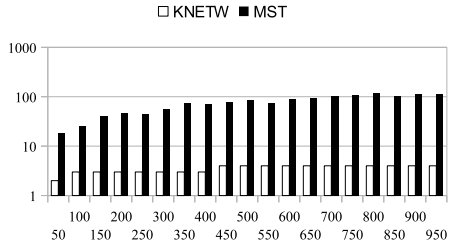


Fig. 7. Max hop-distance from every sensor to the root, $H(\cdot)$.

VII. Conclusions

In this paper we developed various data collection topologies that were based on the location theory notion of centroids and the graph theory notion of balance nodes. We have shown that a centroids based hierarchy provides good approximation factor solutions for energy, transport capacity, and hop-diameter measures, in 2D, and performs asymptotically optimal in 3D for random sensors locations. For arbitrary sensors locations we proved that not much could be done with respect to the energy issue; however we were able to provide a logarithmic height construction that provides logarithmic approximation of the transport capacity objective. Our simulation results verify our theoretical findings and, in fact, suggest that a possible tighter analysis for two

dimensional space may exist. It would also be interesting to investigate the construction, where one of the objectives is an average hop-count between the nodes in the obtained network.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: a survey. *Computer Networks*, 38(4):393–422, 2002.

[2] B. Awerbuch. Optimal distributed algorithms for minimum weight spanning tree, counting, leader election, and related problems. In *ACM STOC 1987*, pages 230–240, 1987.

[3] R. Beier, P. Sanders, and N. Sivadasan. Energy optimal routing in radio networks using geometric data structures. In *ICALP'02*, pages 366–376, 2002.

[4] D. Berend, M. Segal, and H. Shpungin. Energy and lifetime efficient connectivity in wireless ad-hoc networks. *Ad Hoc & Sensor Wireless Networks*, 10(1):61–87, 2010.

[5] G. Calinescu, S. Kapoor, A. Olshevsky, and A. Zelikovsky. Network lifetime and power assignment in ad hoc wireless networks. In *ESA'03*, pages 114–126, 2003.

[6] G. Calinescu, S. Kapoor, and M. Sarwat. Bounded-hops power assignment in ad hoc wireless networks. *Discrete Applied Mathematics*, 154(9):1358–1371, 2006.

[7] A. Chandrakasan, R. Amirtharajah, S. Cho, J. Goodman, G. Konduri, J. Kulik, W. Rabiner, and A. Wang. Design considerations for distributed microsensor systems. In *CICC'99*, pages 279–286, 1999.

[8] J.-H. Chang and L. Tassiulas. Energy conserving routing in wireless ad-hoc networks. In *INFOCOM'00*, pages 22–31, 2000.

[9] A. E. F. Clementi, A. Ferreira, P. Penna, S. Perennes, and R. Silvestri. The minimum range assignment problem on linear radio networks. In *ESA'00*, pages 143–154, 2000.

[10] A. E. F. Clementi, P. Penna, and R. Silvestri. On the power assignment problem in radio networks. *Electronic Colloquium on Computational Complexity*, 7(054), 2000.

[11] A. E. F. Clementi, P. Penna, and R. Silvestri. The power range assignment problem in radio networks on the plane. In *STACS'00*, pages 651–660, 2000.

[12] S. Dolev, M. Segal, and H. Shpungin. Bounded-hop energy-efficient liveness of flocking swarms. *IEEE Transactions on Mobile Computing*, 2012, to appear.

[13] M. Elkin, Y. Lando, Z. Nutov, M. Segal, and H. Shpungin. Novel algorithms for the network lifetime problem in wireless settings. 17(2):397–410, 2011.

[14] S. Funke and S. Soren Laue. Bounded-hop energy-efficient broadcast in low-dimensional metrics via coresets. In *STACS'07*, volume 4393, pages 272–283, 2007.

[15] N. Garg and J. Könemann. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. In *FOCS'98*, pages 300–309, 1998.

[16] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, 2000.

[17] M. M. Halldórsson and P. Mitra. Distributed connectivity of wireless networks. In *PODC*, pages 205–214, 2012.

[18] Ö. D. Incel, A. Ghosh, B. Krishnamachari, and K. Chintalapudi. Fast data collection in tree-based wireless sensor networks. *IEEE Trans. Mob. Comput.*, 11(1):86–99, 2012.

[19] Ö. D. Incel and B. Krishnamachari. Enhancing the data collection rate of tree-based aggregation in wireless sensor networks. In *SECON'08*, pages 569–577, 2008.

[20] A. Jovicic, P. Viswanath, and S. R. Kulkarni. Upper bounds to transport capacity of wireless networks. *IEEE Transactions on Information Theory*, 50(11):2555–2565, 2004.

[21] K. Kalpakis, K. Dasgupta, and P. Namjoshi. Efficient algorithms for maximum lifetime data gathering and aggregation in wireless sensor networks. *Computer Networks Journal*, 42(6):697–716, 2003.

[22] L. M. Kirousis, E. Kranakis, D. Krizanc, and A. Pelc. Power consumption in packet radio networks. *Theoretical Computer Science*, 243(1–2):289–305, 2000.

[23] K.Ramanan and E.Baburaj. Data gathering algorithms for wireless sensor networks: A survey. *IJASUC*, 1, December 2010.

[24] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM*, 57(2):1–32, 2010.

[25] L. Levin, M. Segal, and H. Shpungin. Optimizing performance of ad-hoc networks under energy and scheduling constraints. In *WiOpt'1-*, pages 11–20, 2010.

[26] W. Liang and Y. Liu. Online data gathering for maximizing network lifetime in sensor networks. *IEEE Transactions on Mobile Computing*, 6(1):2–11, 2007.

[27] C. L. Monma and S. Suri. Transitions in geometric minimum spanning trees. *Discrete & Computational Geometry*, 8:265–293, 1992.

[28] A. Orda and B.-A. Yassour. Maximum-lifetime routing algorithms for networks with omnidirectional and directional antennas. In *MobiHoc'05*, pages 426–437, 2005.

[29] K. Pahlavan and A. H. Levesque. *Wireless information networks*. Wiley-Interscience, 1995.

[30] J. Park and S. Sahni. Maximum lifetime broadcasting in wireless networks. *IEEE Transactions on Computers*, 54(9):1081–1090, 2005.

[31] V. Rajendran, K. Obraczka, and J. J. Garcia-Luna-Aceves. Energy-efficient collision-free medium access control for wireless sensor networks. In *SenSys'03*, pages 181–192, 2003.

[32] R. Ramanathan and R. Hain. Topology control of multihop wireless networks using transmit power adjustment. In *INFOCOM'00*, pages 404–413, 2000.

[33] C. Redmond and J. Yukich. Asymptotics for euclidean functionals with power-weighted edges. *Stochastic Processes and their Applications*, 61(2):289 – 304, 1996.

[34] M. Segal. Fast algorithm for multicast and data gathering in wireless networks. *Information Processing Letters*, 2007.

[35] M. Segal and H. Shpungin. On construction of minimum energy k-fault resistant topology. *Ad Hoc Networks*, 7(2):363–373, 2009.

[36] H. Shpungin. Energy efficient online routing in wireless ad hoc networks. In *SECON'11*, pages 64–72, 2011.

[37] H. Shpungin and M. Segal. Low-energy fault-tolerant bounded-hop broadcast in wireless networks. *IEEE/ACM Trans. Netw.*, 17(2):582–590, 2009.

[38] H. Shpungin and M. Segal. Near optimal multicriteria spanner constructions in wireless ad-hoc networks. *IEEE/ACM Transactions on Networking*, 18(6):1963–1976, 2010.

[39] J. Stanford and S. Tongngam. Approximation algorithm for maximum lifetime in wireless sensor networks with data aggregation. In *SNPD'06*, pages 273–277, 2006.

[40] J. M. Steele. Probability and problems in euclidean combinatorial optimization. *Statistical Science*, 8(1):48 – 56, 1993.

[41] L. Su, Y. Gao, Y. Yang, and G. Cao. Towards optimal rate allocation for data aggregation in wireless sensor networks. In *Mobihoc'11*, pages 1–11, 2011.

[42] L. van Hoesel and P. Havinga. A lightweight medium access protocol (lmac) for wireless sensor networks:reducing preamble transmissions and transceiver state switches. In *INSS'04*, pages 205–208, 2004.

[43] Y. Wu, J. A. Stankovic, T. He, and S. Lin. Realistic and efficient multi-channel communications in wireless sensor networks. In *INFOCOM'08*, pages 1193–1201, 2008.

[44] L.-L. Xie and P. R. Kumar. A network information theory for wireless communication: scaling laws and optimal operation. *IEEE Transactions on Information Theory*, 50(5):748–767, 2004.

[45] L.-L. Xie and P. R. Kumar. On the path-loss attenuation regime for positive cost and linear scaling of transport capacity in wireless networks. *IEEE Transactions on Information Theory*, 52(6):2313–2328, 2006.

[46] F. Xue, L.-L. Xie, and P. R. Kumar. The transport capacity of wireless networks over fading channels. *IEEE Transactions on Information Theory*, 51(3):834–0847, 2005.